LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Silke Janitza, Harald Binder, Anne-Laure Boulesteix

Pitfalls of hypothesis tests and model selection on bootstrap samples: causes and consequences in biometrical applications

Technical Report Number 163, 2014
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Pitfalls of hypothesis tests and model selection on bootstrap samples: causes and consequences in biometrical applications

Silke Janitza[1*] Harald Binder[2] Anne-Laure Boulesteix[1]

November 19, 2014

[1] Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany.

[2] Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center Johannes Gutenberg University Mainz, Obere Zahlbacher Str. 69, 55131 Mainz, Germany.

Abstract

The bootstrap method has become a widely used tool applied in diverse areas where results based on asymptotic theory are scarce. It can be applied for example for assessing the variance of a statistic, a quantile of interest or for significance testing by resampling from the null hypothesis. Recently some approaches have been proposed in the biometrical field where hypothesis testing or model selection is performed on a bootstrap sample as if it were the original sample. *P*-values computed from bootstrap samples have been used for example in the statistics and bioinformatics literature for ranking genes with respect to their differential expression, for estimating the variability of *p*-values and for model stability investigations. Procedures which make use of bootstrapped information criteria are often applied in the model stability investigations and model averaging approaches as well as when estimating the error of model selection procedures which involve tuning parameters. From the literature, however, there is evidence that *p*-values and model selection criteria evaluated on bootstrap data sets do not adequately represent what would be obtained on the original data or new data drawn from the overall population. We explain the reasons for this and, through the use of a real data set and simulations, we assess the practical impact on procedures relevant to biometrical applications in cases where it has not yet been studied. Moreover, we investigate the behaviour of subsampling (i.e., drawing from a data set without replacement) as a potential alternative solution to the bootstrap for these procedures.

Keywords: Bootstrap; Bootstrapped information criteria; Bootstrapped *p*-values; Bootstrapped test statistic; Tests on bootstrap samples.

*Corresponding author. Email: janitza@ibe.med.uni-muenchen.de.

1 Introduction

The bootstrap, introduced by Efron (1979), consists of generating a huge number of pseudo-samples from the original data set of interest. In the case of the nonparametric bootstrap (considered in this paper) a pseudo-sample is generated by randomly drawing observations with replacement from the original data. One then typically performs statistical analyses on each bootstrap sample, for instance the computation of an estimator of interest, yielding so-called bootstrapped estimates. Such procedures are becoming more and more widely used, as indicated by the now large number of reference textbooks on the subject (Chernick; 2011; Manly; 2006; Good; 2005; Davison; 1997). Bootstrapped estimates can be used to derive for example the variance of this estimator, a quantile of interest or a confidence interval (Davison; 1997).

In this paper we are interested in the case where a p-value of a standard statistical test (such as, e.g., the Z-test or the likelihood ratio test) takes the role of the estimator which is being bootstrapped. More precisely, we mean p-values that result from statistical tests performed using a bootstrap sample as the data set as if it were the original data set, ignoring that it has actually been drawn with replacement from another sample. For example, a popular bootstrap-based method often applied in biometrical applications investigates the stability of stepwise model selection procedures. This procedure makes use of the bootstrap to generate pseudo-samples, and model selection is performed on each bootstrap sample, where p-values of the likelihood ratio test are used to decide on the inclusion of variables in the model (Chen and George; 1985; Altman and Andersen; 1989; Sauerbrei and Schumacher; 1992).

More concisely, the problem we are addressing in this paper is that p-values computed from bootstrap samples are not valid and cannot be interpreted as p-values: when performing tests on bootstrap samples as if they were realizations from the true unknown distribution, the type I error is increased. This problem has already been reported in the literature for the special case of the likelihood ratio test (Bollen and Stine; 1992) and the χ^2-test (Strobl et al.; 2007) and its consequences for the above mentioned model stability investigations have also been very recently investigated (Rospleszcz et al.; 2014; De Bin et al.; 2014).

It is important to note that although we are interested in a problem also related to bootstrapping and testing, this problem is fundamentally different from obtaining p-values by the so-called bootstrap tests (Efron and Tibshirani; 1994). Bootstrap tests are an alternative to inference based on parametric assumptions when these assumptions are questionable or when such a method simply does not exist. A bootstrap test works roughly as follows: the estimator of interest is computed from a large number of bootstrap samples and the resulting empirical distribution is in some way compared to the null hypothesis. For example, if the null hypothesis states that the parameter of interest equals a certain value, one might look whether this value is within the confidence interval derived from the bootstrap estimates. Bootstrap tests as well as their pitfalls and some potential solutions have been extensively discussed in the literature in recent decades; see Efron and Tibshirani (1994) for an overview. It is important to note that in this paper we are never referring to p-values obtained by such bootstrap tests when we speak of bootstrapped p-values. Instead we are referring to the p-values that are obtained from performing any statistical test (such as, e.g., the Z-test or the likelihood ratio test) using a bootstrap sample as the data set as if it were the original, as outlined in the previous paragraph and which is a completely different approach.

Bootstrapped p-values have been far less investigated than the bootstrap tests outlined in the previous paragraph. However, procedures based on bootstrapped p-values are not uncommon in

the literature, especially in biometrical applications. Besides the example of stability investigations for stepwise model selection procedures mentioned above, p-values computed from bootstrap samples have been used in the statistics and bioinformatics literature for ranking genes with respect to their differential expression (Mukherjee et al.; 2003), for estimating the variability of p-values which one would observe when repeating an experiment multiple times (Boos and Stefanski; 2011) or, in a completely different context, for deciding which variable should be selected for splitting in the recursive partitioning algorithm "random forest", which consists of building decision trees from bootstrap samples (Strobl et al.; 2007).

Similar considerations apply to information criteria like the AIC or BIC. Bootstrapped information criteria are used for example in model stability investigations – as mentioned in the previous paragraph – as well as for model averaging approaches, in which model weights are derived based on bootstrapped AICs (Buckland et al.; 1997; Burnham and Anderson; 2002). Moreover, this issue becomes relevant when estimating the error of model selection procedures which involve tuning parameters. When computed on bootstrap samples, it has been shown that bootstrapped information criteria deviate from information criteria that are computed based on the original sample. In the context of graphical models (Steck and Jaakkola; 2003) and model averaging procedures (Wagenmakers et al.; 2004) this deviation has been shown to lead to a preference for models which are too complex. This issue may also be relevant when using the bootstrap, as an alternative to, say, cross-validation, for estimating the error of a prediction modeling strategy. For a large number of bootstrap samples drawn from the original data set, a prediction model is fit to the bootstrap sample using the considered strategy and is then used to make predictions for the observations which were not included in this bootstrap sample and are thus considered test data. This yields an estimate for the prediction error of the model and the estimates from all bootstrap samples are averaged. Binder and Schumacher (2008) showed that the resulting error estimate is biased in the case where the prediction modeling strategy involves a parameter tuning step based on internal cross-validation. However, as we will show in this paper problems may also occur if the prediction modeling strategy involves a parameter tuning step based on an information criterion like the AIC.

In all these applications it is essential that quantities such as p-values or model selection criteria evaluated on bootstrap samples adequately represent what would be obtained on the original data or new data drawn from the overall population. Several articles, as previously outlined, suggest that this might often not be the case. However, the papers investigating the problems arising from bootstrapping p-values or information criteria are spread over a wide range of very heterogeneous journals which are not often read by biometricians and are to our knowledge not discussed in textbooks. Moreover, they handle very specific cases and a simple general theory to explain the problem is lacking. Further, the practical consequences for biomedical applications are to date largely unknown. The present paper addresses these problems. Its contribution is fourfold: collecting and summarizing empirical and theoretical evidence for this problem from various parts of the literature, providing a new simple theoretical insight into the problem, assessing its practical impact on procedures relevant to biometrical applications in cases where it has not yet been studied, and investigating the behaviour of subsampling (i.e., drawing from a data set without replacement) as a potential alternative solution to the bootstrap.

This paper is structured as follows. In Section 2 we introduce the data set that is used throughout the paper for the investigation of three bootstrap-based procedures. Section 3 both summarizes evidence from the literature and provides new theoretical insight into the problem

of bootstrapping p-values and information criteria. This section also briefly describes the subsampling method which we investigate as a possible alternative to the bootstrap. Using the real data example as well as simulated data, Section 4 addresses practical consequences, and includes comparisons to subsampling. A summary and an outlook are given in Section 5.

2 Data

For our investigations we consider data from the 2007-2008 cycle of the National Health and Nutrition Examination Survey (NHANES) (National Center for Health Statistics; 2012) which is maintained by the Centers for Disease Control and Prevention. NHANES is designed as a series of cross-sectional surveys conducted in the US population. The data are freely available from the institution's homepage or from the Interuniversity Consortium for Political and Social Research. The considered data set comprises a total of $n = 1914$ subjects. For our investigations, we use the level of high-sensitive C-reactive protein (CRP) as the response. The CRP is a plasma protein involved in the acute phase response during inflammatory states (Black et al.; 2004). We included 28 variables in our studies potentially related to the CRP level. A detailed description of these variables is given in the appendix.

Often of interest in studies which include many potentially important variables is which of the variables show evidence for an association with the response. To address this question, p-values obtained from a univariate hypothesis test can be used. In our studies we make use of p-values obtained from the original data and compare the results to those that are obtained when using bootstrapped p-values. Based on these findings we investigate the reliability of univariate variable rankings which are obtained by bootstrapped p-values (as proposed by Mukherjee et al.; 2003, in the context of ranking genes) and show potential pitfalls when relying on such rankings.

Also related to p-values is the question of how much the p-values obtained from an original data set would differ if one were to replicate the same study. In other words we are interested in the variability of the p-values that would be observed if the same study were repeated multiple times. A bootstrap-based procedure has been proposed in which the variability of bootstrapped p-values is used to approximate the true p-value variability (Boos and Stefanski; 2011). Since we have only one replication of the NHANES data it is clear that it is not possible to obtain an estimate for the variability of p-values to which the variance estimate of the bootstrapped p-values can then be compared. For this reason we do not use the NHANES data but simulated data to investigate this issue.

Another issue concerns the choice of an appropriate model describing the association between covariates and the response. Here we first investigate model choice in standard regression models. Information criteria like the AIC or BIC are typically used for choosing a model. With the 28 considered covariates in the NHANES data there are $2^{28} = 268,435,456$ candidate models and, due to computational effort it is not practicable to consider all. Though one usually considers models that include more than one covariate, for ease of illustration in this paper we narrow it down to the 28 models each arising from the inclusion of exactly one of the covariates and investigate which of the models provides the best fit according to the AIC and bootstrapped AIC.

The final examined issue concerns the selection and evaluation of an appropriate model. Here we focus on the special case in which the model selection strategy involves a model selection component that is chosen through an information criterion. A common approach is to use a bootstrap sample for model selection and to use the observations that are not part of the bootstrap

sample for model evaluation. We apply the model selection strategy on the complete NHANES data to find an appropriate boosting model and then compare the results to those that are obtained when we apply the model selection strategy to a bootstrap sample, a common strategy to assess the accuracy of the chosen model.

3 Methods

In this section we summarize evidence from the literature and provide new theoretical insight into the problem of tests performed on bootstrap samples (Section 3.1), including the related issue of information criteria – such as the AIC – computed from bootstrap samples (Section 3.2). We also briefly describe the subsampling method investigated in this paper as a possible alternative to the bootstrap (Section 3.3).

3.1 Bootstrapping p-values

3.1.1 Z-test

Let $\boldsymbol{x}^\top = (x_1, \ldots, x_n)$ be realizations drawn from $N(\mu, \sigma^2)$ and let $\hat{F}$ denote the corresponding empirical distribution. The test statistic for testing the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_1 : \mu \neq \mu_0$ is given by $Z = \sqrt{n}(\bar{x} - \mu_0)/\sigma$, with $\bar{x}$ denoting the sample mean. Then Z follows a normal distribution with $\mathrm{E}(Z) = \sqrt{n}(\mu - \mu_0)/\sigma$ and $\mathrm{Var}(Z) = 1$.

Now let $\boldsymbol{x}^{*\top} = (x_1^*, \ldots, x_n^*)$ denote the realizations of a bootstrap sample that was drawn from $\hat{F}$ with replacement. The bootstrapped test statistic from a Z-test with hypotheses $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ is defined as

$$Z^* = \sqrt{n}\frac{\bar{x}^* - \mu_0}{\sigma}, \tag{1}$$

with $\bar{x}^* = \frac{1}{n}\sum_{i=1}^n x_i^*$ and known σ. Z^* does not follow the same distribution as Z, as stated in Theorem 1.

Theorem 1

Let the bootstrapped statistic for a Z-test with $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ be defined as in Eq. (1). The expectation of this bootstrapped Z-test statistic Z^* is $\mathrm{E}(Z^*) = \mathrm{E}(Z)$, while the variance of Z^* is $\mathrm{Var}(Z^*) = 2$.

Proof

We derive

$$\mathrm{E}(Z^*) = \mathrm{E}(\mathrm{E}(Z^*|\hat{F})) = \mathrm{E}(Z).$$

The variance of Z^* can be split into two parts,

$$\mathrm{Var}(Z^*) = \mathrm{Var}(\mathrm{E}(Z^*|\hat{F})) + \mathrm{E}(\mathrm{Var}(Z^*|\hat{F})). \tag{2}$$

The first term reduces to

$$\mathrm{Var}(\mathrm{E}(Z^*|\hat{F})) = \mathrm{Var}(Z|\hat{F}) = 1. \tag{3}$$

As far as the second term in (2) is concerned, the basic assumption underlying bootstrap estimation of the variance, which can be easily shown in the present simple case (Davison; 1997), is that

Hypotheses	Sign. threshold	Type I error for Z	Type I error for Z^*
$H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$	$z_{0.95} = 1.64$	0.10	0.24
(two-sided test)	$z_{0.975} = 1.96$	0.05	0.17
	$z_{0.995} = 2.58$	0.01	0.07
$H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$	$z_{0.90} = 1.28$	0.10	0.18
(one-sided test)	$z_{0.95} = 1.64$	0.05	0.12
	$z_{0.99} = 2.33$	0.01	0.05

Table 1: Type I error when performing two-sided and one-sided upper Z-tests with pre-defined significance thresholds for test statistic $Z = \sqrt{n}(\frac{1}{n}\sum x_i - \mu_0)/\sigma$ with $x_1, \ldots, x_n \overset{iid}{\sim} N(\mu, \sigma)$, and for a bootstrapped test statistic $Z^* = \sqrt{n}(\frac{1}{n}\sum x_i^* - \mu_0)/\sigma$.

$\mathrm{Var}(Z^*|\hat{F})$ approximates $\mathrm{Var}(Z)$. Using this result one obtains for the second term

$$\mathrm{E}(\mathrm{Var}(Z^*|\hat{F})) = \mathrm{E}(\mathrm{Var}(Z)) = 1. \tag{4}$$

Summing (3) and (4), Eq. (2) results in $\mathrm{Var}(Z^*) = 2$.

□

According to Theorem 1, the bootstrapped statistic Z^* has twice the variance of Z. Thus under the null hypothesis that $H_0 : \mu = \mu_0$ (or $H_0 : \mu \leq \mu_0$; $H_0 : \mu \geq \mu_0$ for one-sided tests), the bootstrapped statistic Z^* does not follow the standard normal distribution (see also the appendix for empirical results). If incorrectly assuming a standard normal distribution for Z^* under H_0, the derived p-values are biased: using the significance threshold $z_{1-\frac{\alpha}{2}}$, the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution, the type I error is $2 \cdot (1 - \Phi(\frac{1}{\sqrt{2}} z_{1-\frac{\alpha}{2}}))$, where Φ is the standard normal distribution function. For a one-sided lower (upper) test with null hypothesis $H_0 : \mu \geq \mu_0$ ($H_0 : \mu \leq \mu_0$), the significance threshold z_α ($z_{1-\alpha}$) is used and the type I error is $\Phi(\frac{1}{\sqrt{2}} z_\alpha)$ (and $1-\Phi(\frac{1}{\sqrt{2}} z_{1-\alpha})$, respectively). Table 1 shows examples for the type I error when performing Z-tests for test statistics Z and Z^*. It can be seen that the type I error is substantially increased when (incorrectly) assuming that the bootstrapped test statistic Z^* follows a standard normal distribution.

3.1.2 Likelihood Ratio Test

The likelihood ratio (LR) test is used for example when comparing the fit of two nested models, where one model contains restrictions that are not imposed in the other. The likelihood of the restricted model, called the submodel in the following, is termed L_0, while L_1 corresponds to the likelihood of the unrestricted model. The test statistic for the LR test is defined as twice the difference in log-likelihoods:

$$T = -2(\log(L_0) - \log(L_1)). \tag{5}$$

The test statistic T asymptotically follows a non-central χ^2-distribution with df degrees of freedom, which is calculated as the difference in degrees of freedom of the two models, and with non-centrality parameter κ. The asymptotic expectation of the test statistic is given by $\mathrm{AE}(T) = df + \kappa$ and the asymptotic variance is $\mathrm{AVar}(T) = 2df + 4\kappa$. Under the null hypothesis which states that the submodel is true, the non-centrality parameter is zero and thus T asymptotically follows a central $\chi^2(df)$-distribution and has asymptotic expectation $\mathrm{AE}(T) = df$ and asymptotic variance $\mathrm{AVar}(T) = 2df$.

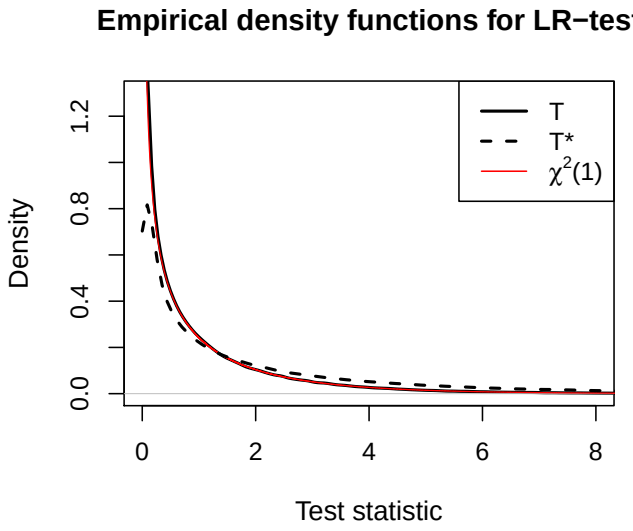

Figure 1: Empirical density functions for test statistics T (solid black line) and T^* (dashed black line) of the LR test with 1 degree of freedom. The density of the $\chi^2(1)$-distribution is indicated by the red line.

Bollen and Stine (1992) gave an approximation for the asymptotic expectation of the test statistic T^* that is derived from the bootstrap sample $\boldsymbol{x}^*$. They report it as being twice as large as the asymptotic expectation of T in the original sample. They also report the asymptotic variance of T^* to be larger than the asymptotic variance of T. These theoretical results are in line with our empirical results from a simulation study for which we used a very large sample size of $n = 100000$. For this study, we draw predictor values $X_i \sim N(0, 1)$ and, independently of the predictor values, we draw response values $Y_i \sim N(0, 1)$ for observations $i = 1, \ldots, n$. Subsequently a bootstrap sample is drawn from this original sample. A LR test with one degree of freedom is performed on the original sample and on the bootstrap sample to test if the linear regression model which includes predictor X gives a better model fit than the intercept model. From this test we obtain test statistics T for the original sample and T^* for the bootstrap sample. The data generation and computation of the test statistics are repeated 500000 times in order to obtain empirical distributions of T and T^*.

Figure 1 shows the empirical density functions of T and T^*. The distribution of T approximates the χ^2-distribution with 1 degree of freedom very well (the respective lines in Figure 1 coincide), which indicates that the number of observations was chosen large enough. It is remarkable that the distribution of T^* so noticeably deviates from the χ^2-distribution. It has a much higher variability, so the probability mass in the tail is larger than that of T. To quantify the discrepancy between the empirical distributions of T and T^* we compute the empirical expectation (sample mean) and empirical variance of T and T^*. While the empirical expected value of T is, at 0.9977, very close to the true asymptotic expectation of 1, the empirical value of T^* is 2.0005 and is thus approximately twice as large, as predicted by the approximation provided by Bollen and Stine (1992). The empirical variance of T is, at 1.9819, also close to the theoretical approximate variance of 2. The variance of T^* in contrast is, with a value of 8.0140, higher by a factor of 4. From these results it is obvious that the type I error is increased when performing a LR test on

bootstrap samples using the critical values from a χ^2-distribution.

3.2 Bootstrapping Information Criteria for Model Building

Information criteria are often used for the comparison of non-nested models. These measures compare models based on their goodness-of-fit to the data while penalizing the complexity of the model (see also Burnham and Anderson; 2002). Akaike's information criterion (AIC) is a widely used measure for model selection. It is defined as

$$\text{AIC} = -2\log(L) + 2p, \tag{6}$$

where L denotes the likelihood and p denotes the number of parameters included in the model. It has been shown that minimizing the AIC is approximately equivalent to minimizing the expected Kullback-Leibler distance between the true and the estimated density (Akaike; 1973).

The bootstrapped AIC is given by

$$\text{AIC}^* = -2\log(L^*) + 2p, \tag{7}$$

with L^* denoting the likelihood computed for a model that was fit on a bootstrap sample. To prove that this bootstrapped AIC is not a good approximation of the AIC defined in (6), we will compare two nested models using the AIC, with the models differing in the inclusion of only one parameter (similar considerations can be made in the case of nested models differing by the inclusion of more than one parameter). If AIC_1 denotes the AIC of the unrestricted model that includes p parameters and AIC_0 denotes the AIC of the submodel that includes $p-1$ parameters, then the LRT on one degree of freedom can be expressed in terms of AIC_0 and AIC_1 as follows (cf. Chapter 6.9.3 in Burnham and Anderson; 2002):

$$\text{LRT} = \text{AIC}_0 - \text{AIC}_1 + 2. \tag{8}$$

From Eq. (8) we see that if both models fit the data equally well according to the AIC (i.e., $\text{AIC}_0 = \text{AIC}_1$), we have LRT = 2. Further, the unrestricted model is chosen over the submodel if its AIC is smaller, corresponding to $\text{AIC}_0 - \text{AIC}_1 > 0$ and, according to Eq. (8), LRT > 2. In contrast, the submodel is chosen if $\text{AIC}_0 - \text{AIC}_1 < 0$, corresponding to LRT < 2. These considerations show that in the case of two nested models one can also use the value of the LRT to decide which of the models is better in terms of the AIC; values for the LRT below 2 indicate the superiority of the submodel, values above 2 indicate that the unrestricted model is better, and both models are considered equally good if the LRT takes the value 2. As shown in 3.1.2, bootstrapped LR test statistics systematically deviate from LR test statistics derived from the original data. Due to the correspondence between the LRT and the AIC in the specific setting of nested models it follows that bootstrapped information criteria like the AIC are thus not valid either.

These considerations were also made by Wagenmakers et al. (2004). In the context of graphical models, Steck and Jaakkola (2003) proved that bootstrapped information criteria systematically deviate from information criteria derived from original samples.

3.3 Subsampling as an Alternative to the Bootstrap

The subsampling procedure, also known as delete-d jackknife (Wu; 1986), is closely related to the bootstrap, but in contrast to the bootstrap a subsample is created by drawing m observations, with $m < n$, without replacement from the original sample. The optimal choice of the parameter m is delicate and is not treated here (see, e.g., Davison et al.; 2003; Bickel and Sakov; 2005). In our studies we chose m as the value $0.632n$, which corresponds to the expected number of unique observations in a bootstrap sample, in order to have on average the same number of unique observations in subsamples and bootstrap samples. The subsampling technique has been investigated in the literature and also contrasted to the bootstrap (Shao and Wu; 1989; Politis and Romano; 1994; Politis et al.; 1999; Hartigan; 1969). It shows asymptotic consistency in cases where the bootstrap fails (Davison et al.; 2003; Chernick; 2011). In particular the type I error is not increased for test statistics computed on subsamples. For this reason Strobl et al. (2007) recommended using subsampling instead of bootstrapping in the random forest algorithm.

4 Results

4.1 Bootstrapping p-values

In the following the aim is to investigate the strength of the association between each of the 28 covariates and the response (i.e., the level of high-sensitive C-reactive protein). For this purpose we apply a univariate test for each covariate and use the p-value as a measure for the strength of association.

We obtained p-values in the range $[0, 0.6]$ with 17 p-values lying below 0.05, which is strong evidence for associations in the data. Besides computing p-values based on the original NHANES sample, we also investigated bootstrapped p-values. To obtain stable results we computed p-values for $B = 10000$ bootstrap samples of the NHANES data and computed the median value. Figure 2 (left) shows the obtained median bootstrapped p-values for each of the 28 covariates plotted against the p-value that was obtained for the original sample. As one can clearly see, bootstrapped p-values systematically deviate from the p-values obtained for the original sample. More precisely, in many cases bootstrapped p-values are considerably too small, suggesting stronger evidence for the association between the covariates and the CRP level than evidenced by the original data.

Figure 3 (left) shows the relative frequency of significant associations in the $B = 10000$ bootstrap samples (at nominal α-level of 0.05). On average there were 18.4 significant associations, while in the original data 17 of the 28 associations were significant. As expected from theory there are more significant associations when using bootstrapped p-values.

From the p-values computed based on the original NHANES sample, we have observed strong evidence for associations between the CRP level and many of the covariates; however, it is also interesting to investigate the results for a data set with less evidence for associations. Thus we repeated the calculations using modified versions of the NHANES data in the extreme case of no associations between the covariates and the response. To obtain such modified versions of the NHANES data we randomly permuted the response variable to break the association between all of the 28 covariates and the response. This was repeated 1000 times to obtain a total of 1000 data sets in which no associations are present between the covariates and the response. For the sake of clarity we show the results only for the first 10 modified NHANES data sets (right panel in Figure 2). Note that since we have 10 data sets, now 10×28 points are plotted.

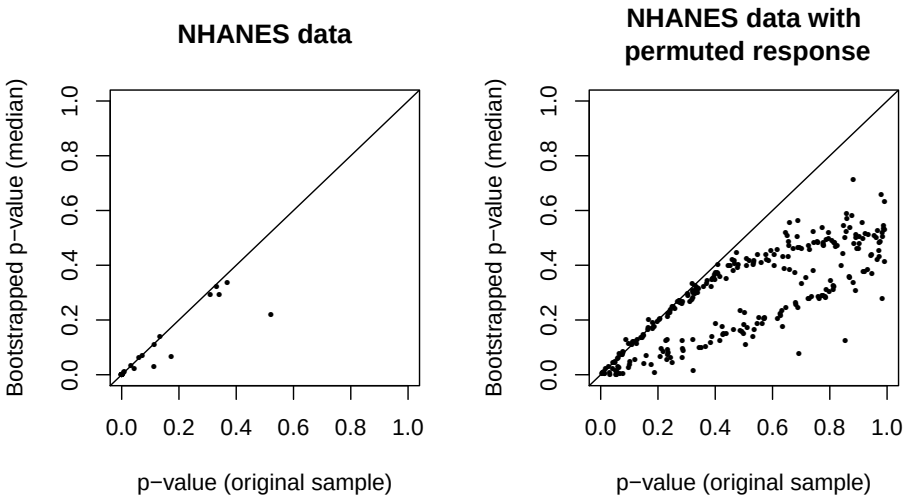

Figure 2: p-values obtained for testing the association between CRP level and each of the 28 covariates in the NHANES sample plotted against the median p-value of $B = 10000$ bootstrapped p-values. Points lying on the diagonal line would indicate agreement between p-values derived on the original NHANES data and bootstrapped p-values. Left: Results obtained for the unmodified NHANES data. Right: Results obtained for the 10 modified NHANES data sets in which there are no associations between covariates and the CRP level (via permuting values for CRP level).

Again the median bootstrapped p-values are substantially smaller than p-values for the original samples. As expected, p-values computed for the modified data sets are scattered in the interval $[0, 1]$; on average in the 1000 permuted data sets 1.36 of the 28 p-values take a value below 0.05. While p-values for the original samples fully cover the interval $[0, 1]$, bootstrapped p-values only take values in the lower range $[0, 0.6]$ (see Figure 2), suggesting stronger associations than are actually present in the data and underlining the fact that bootstrapped p-values are not uniformly distributed on $[0, 1]$ under the null hypothesis. This can also be seen when computing the number of significant associations from bootstrapped p-values. The number of significant associations is strongly overestimated, as seen in the right panel of Figure 3. While there are 1.36 significant associations on average in the original samples, the average number (taken over all $1000 \times B$ bootstrap samples) of significant associations according to bootstrapped p-values is 6.12.

We performed the same computations using subsamples instead of bootstrap samples, with results shown in Figures 4 and 5. From theory it is clear that p-values obtained from subsamples systematically deviate from p-values obtained for the original sample due to the smaller sample size and the decreased power to detect associations in subsamples: this is clearly seen in Figure 4. On average 14.7 of the 28 covariates were significantly associated with the CRP level in subsamples compared to 17 significant associations in the original sample (see also the left panel of Figure 5).

In the case where no associations exist – the NHANES data with permuted response – a comparable number of significant findings can be observed in subsamples and in the 1000 original samples: there were on average 1.40 significant associations in subsamples compared to 1.36 significant findings in the 1000 original samples. This is in line with the fact that tests performed on subsamples – in contrast to tests performed on bootstrap samples – do not have an increased type I error (see, e.g, Sauerbrei et al.; 2011). Accordingly, p-values derived on subsamples may

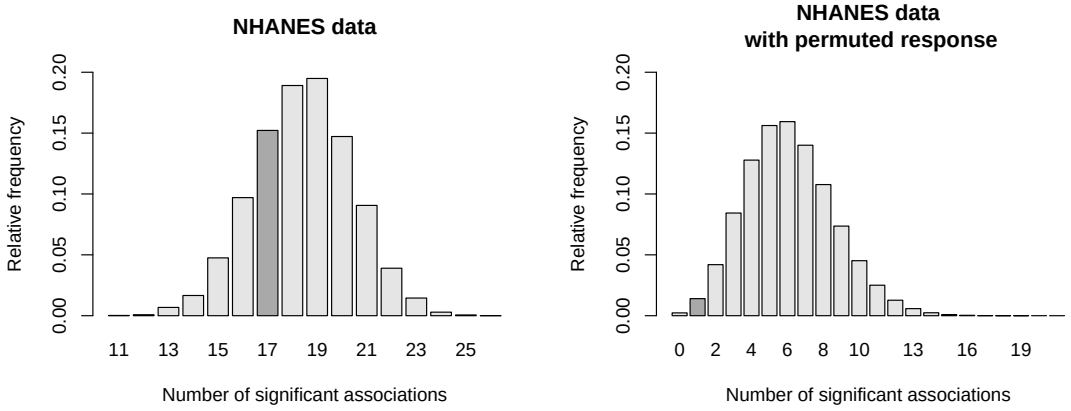

Figure 3: Relative frequency of bootstrap samples (of 10000) with specified number of significant results when univariately testing the association between CRP level and 28 covariates. The dark gray bar indicates the number of significant associations in the unmodified NHANES data (left) and in the NHANES data in which there are no associations between covariates and the CRP level (via permuting the response, right).

be used for testing a specific hypothesis.

Here we have seen that bootstrapped p-values systematically deviate from those that are obtained for the original sample. In the next section we investigate the consequences when using bootstrapped p-values for ranking variables. Such an approach has been proposed by Mukherjee et al. (2003) for ranking genes with respect to their differential expression.

4.2 Bootstrapped p-values for Variable Ranking

Previous studies by Bollen and Stine (1992) and Rospleszcz et al. (2014) showed that the discrepancy in bootstrapped p-values obtained from a LR test is more pronounced for categorical predictor variables with many categories. More precisely, the increase in type I error in the LR test has been shown to depend on the degrees of freedom of the test. Since a categorical predictor with m categories is represented by $m-1$ dummy variables, a LR test that tests for the significance between this predictor and the CRP level has $m-1$ degrees of freedom. A metric variable in contrast is represented by one parameter and the corresponding LR test has one degree of freedom. Accordingly, when computing p-values based on bootstrap samples, the p-values are more greatly underestimated for categorical predictors with many categories. A similar mechanism has been observed by Strobl et al. (2007) in the context of the χ^2-test. This issue and its consequences for model building procedures and the random forest method have been extensively investigated by Rospleszcz et al. (2014) and Strobl et al. (2007), respectively. Here we show that this issue also impacts the results obtained for a variable ranking approach that was proposed by Mukherjee et al. (2003) in the context of gene expression studies. This approach consists of computing the p-values for a large number of bootstrap samples, obtaining the median p-value for all considered genes, and sorting the genes by the median p-value. In the following we apply this approach to the NHANES data to obtain a ranking of the 28 considered variables.

Figure 6 shows the variable rankings for the unmodified NHANES data. The upper left panel corresponds to the rankings according to the p-values from the original NHANES sample and the

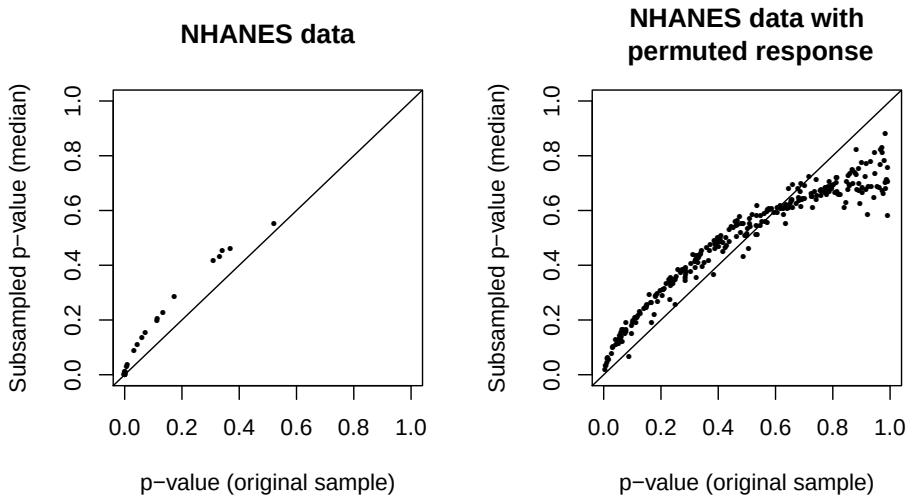

Figure 4: p-values obtained for testing the association between CRP level and each of the 28 covariates in the NHANES sample plotted against the median p-value of $B = 10000$ p-values computed on subsamples. Points lying on the diagonal line would indicate agreement between p-values derived on the original NHANES data and p-values derived on subsamples. Left: Results obtained for the unmodified NHANES data. Right: Results obtained for the modified NHANES data in which there are no associations between covariates and the CRP level (via permuting values for CRP level).

right upper panel corresponds to rankings by the median bootstrapped p-values (i.e., the median of $B = 10000$ bootstrapped p-values). In addition, results are shown when using the median p-value obtained from $B = 10000$ subsamples (lower panel).

On the whole, the rankings are similar for the unmodified NHANES data, especially among those variables with strong evidence for association. However, close inspection reveals some differences between the rankings based on the original sample and those based on bootstrap samples. More precisely, we observed the phenomenon – described in Rospleszcz et al. (2014) – that categorical predictors with many categories obtain systematically smaller bootstrapped p-values than metric predictors or categorical predictors with fewer categories. In our variable ranking this can be seen from the fact that variables with many categories gain ranking positions closer to the top when ranked by the median bootstrapped p-value. Table 2 shows the ranking positions for each variable separately for variables of different scale. There are numerousness cases in which categorical variables with four or more categories gain ranking positions closer to the top when ranked by bootstrapped p-values. Conversely, the binary and metric variables are located at positions at the bottom of the ranking when the ranking is according to bootstrapped p-values.

In contrast, when using subsamples there are only minor differences in the ranking, with seemingly no effect of a variable's scale on its ranking position. This is also in line with the results presented in Rospleszcz et al. (2014) who investigated the use of subsampling as an alternative to bootstrap for a model building procedure.

The observed mechanisms are even more extreme for the modified NHANES data sets where the response variable had been permuted (see Table A2, which shows the result for the first modified data set). For the modified data sets there are very large differences in the variable ranking – with variables with many categories ranked at top positions and binary or metric variables at much

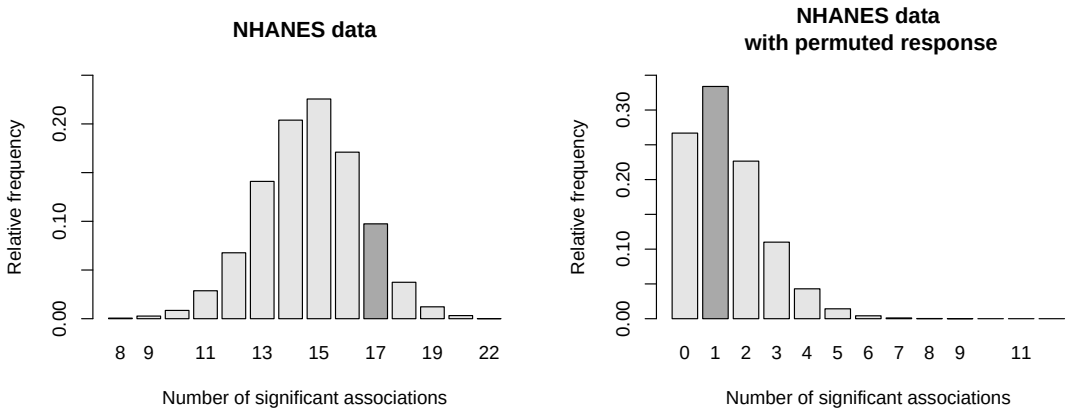

Figure 5: Relative frequency of subsamples (of 10000) with specified number of significant results when univariately testing the association between CRP level and 28 covariates. The dark gray bars indicate the number of significant associations in the unmodified NHANES data (left) and in the NHANES data in which there are no associations between covariates and the CRP level (via permuting the response, right).

lower positions – when p-values are derived from bootstrap samples.

To conclude, our studies show that, though resampling procedures might be promising methods for obtaining stable variable ranking lists, bootstrapped p-values should not be compared with significance thresholds for making decisions on the significance of variables. In particular, care needs to be taken when the interest lies in ranking variables of different scales, which often occurs in epidemiological studies. An example of further relevance is gene ranking when single nucleotide polymorphisms are considered, which for some genes are represented by a categorical variable with three categories but for others only two categories. Moreover associations between genes and a phenotype are usually weak or non-existent, which is expected to be especially problematic as suggested by our results. Thus in settings including categorical predictors bootstrapped p-values should not be applied for obtaining ranking lists.

Subsampling may be a reasonable alternative to the bootstrap for variable ranking: in our studies there were only minor differences between the ranking lists – as determined by sorting variables by p-value – obtained from the original sample and from subsamples. This might indicate that in the considered NHANES data set there are not many influential points that have a large impact on the results, but more research is needed on this topic. We conclude from these results that subsampling should be preferred over bootstrapping for obtaining variable rankings if variables are of different scales. We have to note, however, that in settings with very small sample sizes – for which the ranking approach was originally proposed (Mukherjee et al.; 2003) – subsampling from a data set that consists of only a few observations may not be advisable.

4.3 Bootstrapped p-values for Assessing the Variability of p-values

In Section 4.1 we saw that bootstrapped p-values systematically deviate from p-values from the original data and lead to more significant findings than are supported by the data. The bias in bootstrapped p-values is likely to have impact on other measures that are computed using bootstrapped p-values, such as the variance of p-values. Recently it has been proposed to compute

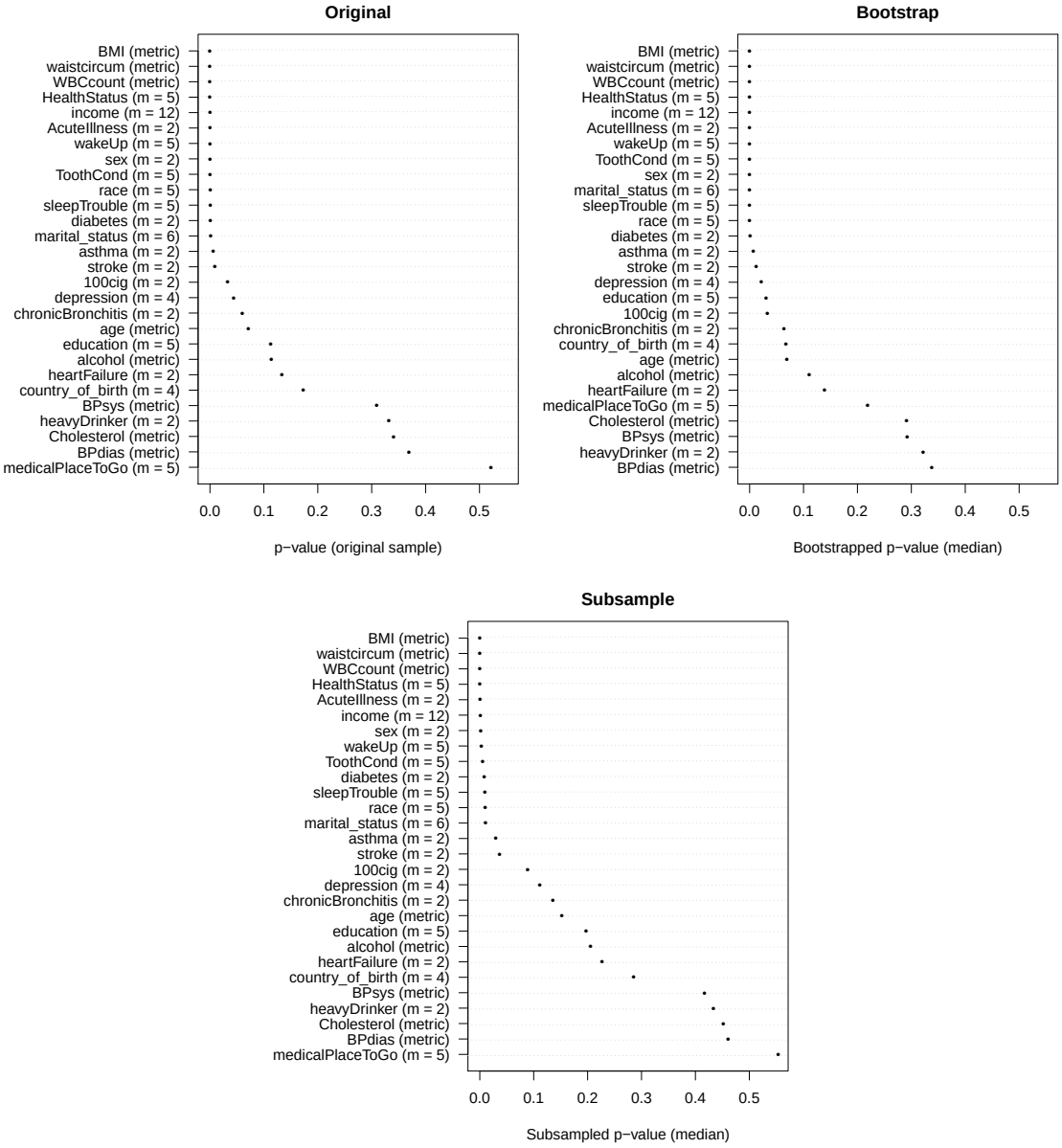

Figure 6: Variable ranking by p-values obtained for the original NHANES sample (upper left) and by the p-value obtained from the median over $B = 10000$ bootstrapped p-values (upper right) and the median p-value from subsamples (lower). The parameter m denotes the number of levels of a categorical predictor variable.

the variance of p-values, or preferably the variance of -$\log_{10}$(p-value) (Boos and Stefanski; 2011). The question arises of whether the variance of bootstrapped p-values can be used to approximate the variability of p-values that would be observed if we repeatedly performed the same experiment. To investigate this issue we performed simulation studies, which allow us to draw multiple times from the true distribution F.

In the first part of our simulation studies we independently drew $n = 1000$ observations from $N(0, 1)$ and tested $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. One bootstrap sample was generated by drawing from this data with replacement. A Z-test (as described in Section 3.1.1) was performed separately

Scale	Variable	Original rank	Bootstrap rank	(diff.)	Subsample rank	(diff.)
metric or $m = 2$	BMI	1	1	(0)	1	(0)
	waistcircum	2	2	(0)	2	(0)
	WBCcount	3	3	(0)	3	(0)
	AcuteIllness	6	6	(0)	5	(+1)
	sex	8	9	(−1)	7	(+1)
	diabetes	12	13	(−1)	10	(+2)
	asthma	14	14	(0)	14	(0)
	stroke	15	15	(0)	15	(0)
	100cig	16	18	(−2)	16	(0)
	chronicBronchitis	18	19	(−1)	18	(0)
	age	19	21	(−2)	19	(0)
	alcohol	21	22	(−1)	21	(0)
	heartFailure	22	23	(−1)	22	(0)
	BPsys	24	26	(−2)	24	(0)
	heavyDrinker	25	27	(−2)	25	(0)
	Cholesterol	26	25	(+1)	26	(0)
	BPdias	27	28	(−1)	27	(0)
$m = 4$	depression	17	16	(+1)	17	(0)
	country_of_birth	23	20	(+3)	23	(0)
$m = 5$	HealthStatus	4	4	(0)	4	(0)
	wakeUp	7	7	(0)	8	(−1)
	ToothCond	9	8	(+1)	9	(0)
	race	10	12	(−2)	12	(−2)
	sleepTrouble	11	11	(0)	11	(0)
	education	20	17	(+3)	20	(0)
$m = 6$	medicalPlaceToGo	28	24	(+4)	28	(0)
	marital_status	13	10	(+3)	13	(0)
$m = 12$	income	5	5	(0)	6	(−1)

Table 2: Variable ranking for the unmodified NHANES data. Variable rankings are obtained from p-values obtained for the original NHANES sample ("Original rank"), from the median bootstrapped p-value ("Bootstrap rank"), and from the median p-value from subsamples ("Subsample rank"). The difference to the "Original rank" is given in brackets for each variable. The parameter m denotes the number of levels of a categorical predictor variable.

for the original sample and the bootstrap sample. This process was repeated 5000 times. The same analysis was done for the case where the n observations came from $N(0.08, 1)$.

In the second part of our simulation studies p metric predictor variables $x_{i1}, \ldots, x_{ip}$ were independently drawn for $i = 1, \ldots, 1000$ from a multivariate normal distribution with expected value $\boldsymbol{\mu} = (0, \ldots, 0)^\top \in \mathbb{R}^p$ and variance $\boldsymbol{I}_p$, corresponding to the identity matrix of dimension p. The response variable Y_i was generated according to the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

with $\epsilon_i \sim N(0, 1)$. The global null hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ states that none of the p predictors is associated with the response, and the alternative hypothesis is that at least one of the coefficients is associated, that is $H_1 : \beta_j \neq 0$ for at least one $j \in \{1, \ldots, p\}$. The corresponding LR test compares the likelihood of the submodel L_0 which contains only the intercept, to the likelihood L_1 of the model which contains all predictor variables. If the null hypothesis is true the LR test statistic (5) follows a central χ^2-distribution with p degrees of freedom. In our simulations

all beta coefficients were set to the value zero (i.e., the null hypothesis is true). As before, p-values were derived from 5000 original samples and from 5000 bootstrap samples. An additional analysis was performed in which the alternative hypothesis is true. For this simulation all beta coefficients were set to 0.02.

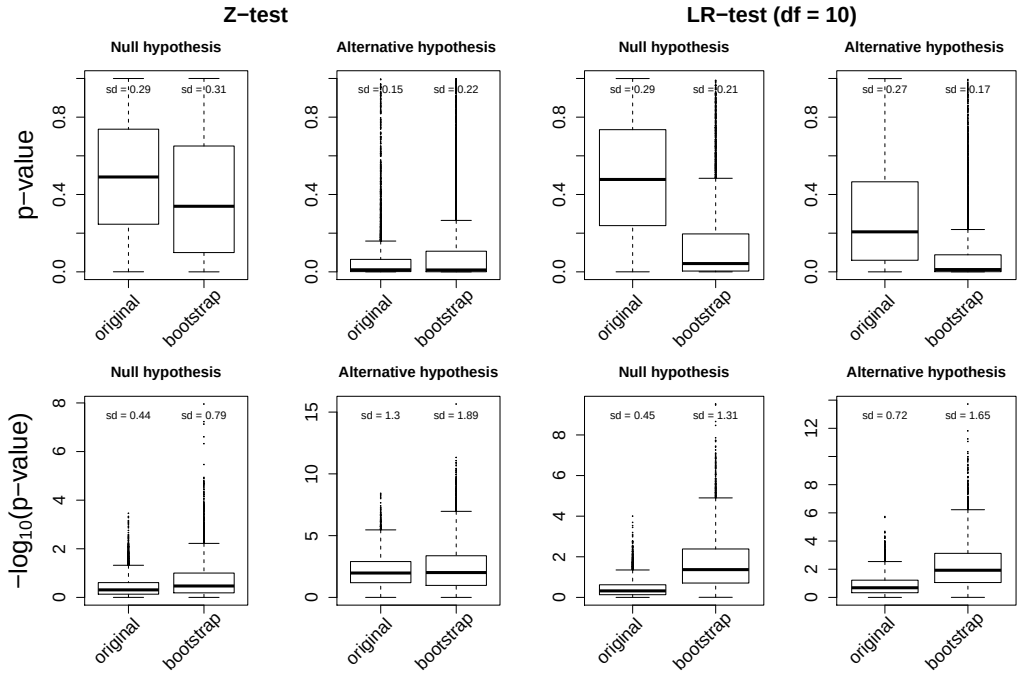

Figure 7: p-value distribution for the Z-test (left two columns) and the LR test with 10 degrees of freedom (right two columns) computed on 5000 original samples (left boxplot) and 5000 bootstrap samples (right boxplot) under the null hypothesis and under the alternative hypothesis. The standard deviation (sd) as a measure for p-value variability is given above each boxplot.

Figure 7 shows the distributions of p-values and standard deviations for the first (Z-test) and the second simulation study (LR test): it is clear that the p-value distribution derived from bootstrap samples is not a good approximation of the p-value distribution that is obtained for original samples. The standard deviations of the p-value (or -$\log_{10}$(p-value)) computed on bootstrap samples do not reflect the true p-value variability in our studies, neither under the null hypothesis nor under the alternative hypothesis. Subsampling is not a reasonable alternative here if more than estimation of the p-value distribution under the null hypothesis and type I error control is wanted. Figure 8 shows the corresponding results for the Z-test and the LR test when the test is performed on subsamples. One can see that tests performed on subsamples preserve the α-level but one obtains higher p-values under the alternative hypothesis, due to the decreased statistical power. Tests performed on subsamples thus do not reflect the p-value variability in original samples either.

4.4 Bootstrapping Information Criteria

In Section 4.1 we computed p-values to assess the strength of association between CRP level and each of the 28 covariates. For this purpose we fit one model to each of the covariates and tested if including the covariate significantly improves the model fit compared to the intercept model. Now

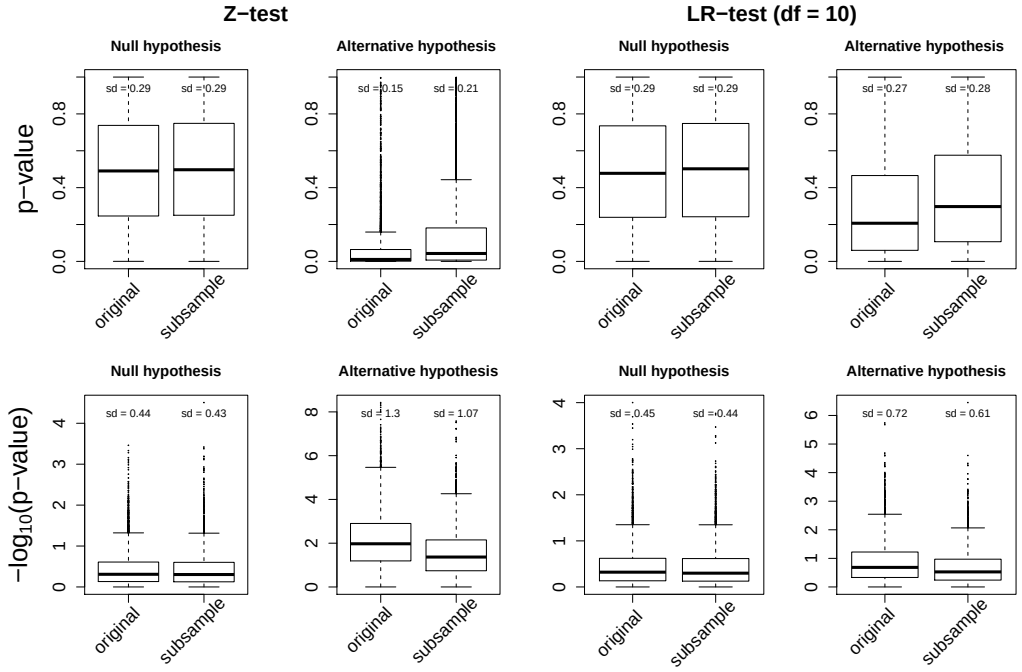

Figure 8: p-value distribution for the Z-test (left two columns) and the LR test with 10 degrees of freedom (right two columns) computed on 5000 original samples (left boxplot) and 5000 subsamples (right boxplot) under the null hypothesis and under the alternative hypothesis. The standard deviation (sd) as a measure for p-value variability is given above each boxplot.

we are interested in the goodness-of-fit of the 28 models and especially in the question of which models provide the best fit. Since parameters are of different scale, one cannot directly compare the likelihood of the models. In such cases information criteria like the AIC are often used. We derived AIC values for models fit on the original NHANES sample as well as for models fit on bootstrap samples. The left panel of Figure 9 shows the ranking of models by AIC value obtained for the original sample. Bootstrapped AIC values were computed for $B = 10000$ bootstrap samples and an average AIC value was computed. The right panel of Figure 9 shows the ranking by this average bootstrapped AIC. While the top and the bottom of the ranking lists are nearly identical, a number of differences can be observed in the middle: the model which includes $k = 5$ parameters coding marital status is ranked at the 12th position based on the original NHANES sample, while based on bootstrap samples it is ranked 9th. Conversely, the model which includes the variable sex ($k = 1$) was ranked 9th based on the original sample but only 12th when AICs were derived from bootstrap samples. Considerable differences in the ranking position can also be observed for the model which includes educational background ($k = 4$). For the original sample this model was ranked only 22nd, while for bootstrap samples it is ranked 17th. Overall, when looking at both rankings, one can see that models which include more parameters seem to obtain higher rankings when ranked by bootstrapped AICs. This applies for the models based on the covariates wakeUp, sleepTrouble, marital_status, depression, education, or country_of_birth. Models which include only one parameter (in addition to the intercept) have lower rankings for bootstrapped AICs (for covariates: sex, acuteIllness, 100cig, chronicBronchitis, age, alcohol, heartFailure, BPsys, heavyDrinker). There are only two exceptions where it is reverse (cholesterol and race). These results strongly suggest that there is a preferential selection of more complex models – i.e., those

that include more parameters – when using bootstrapped AICs.

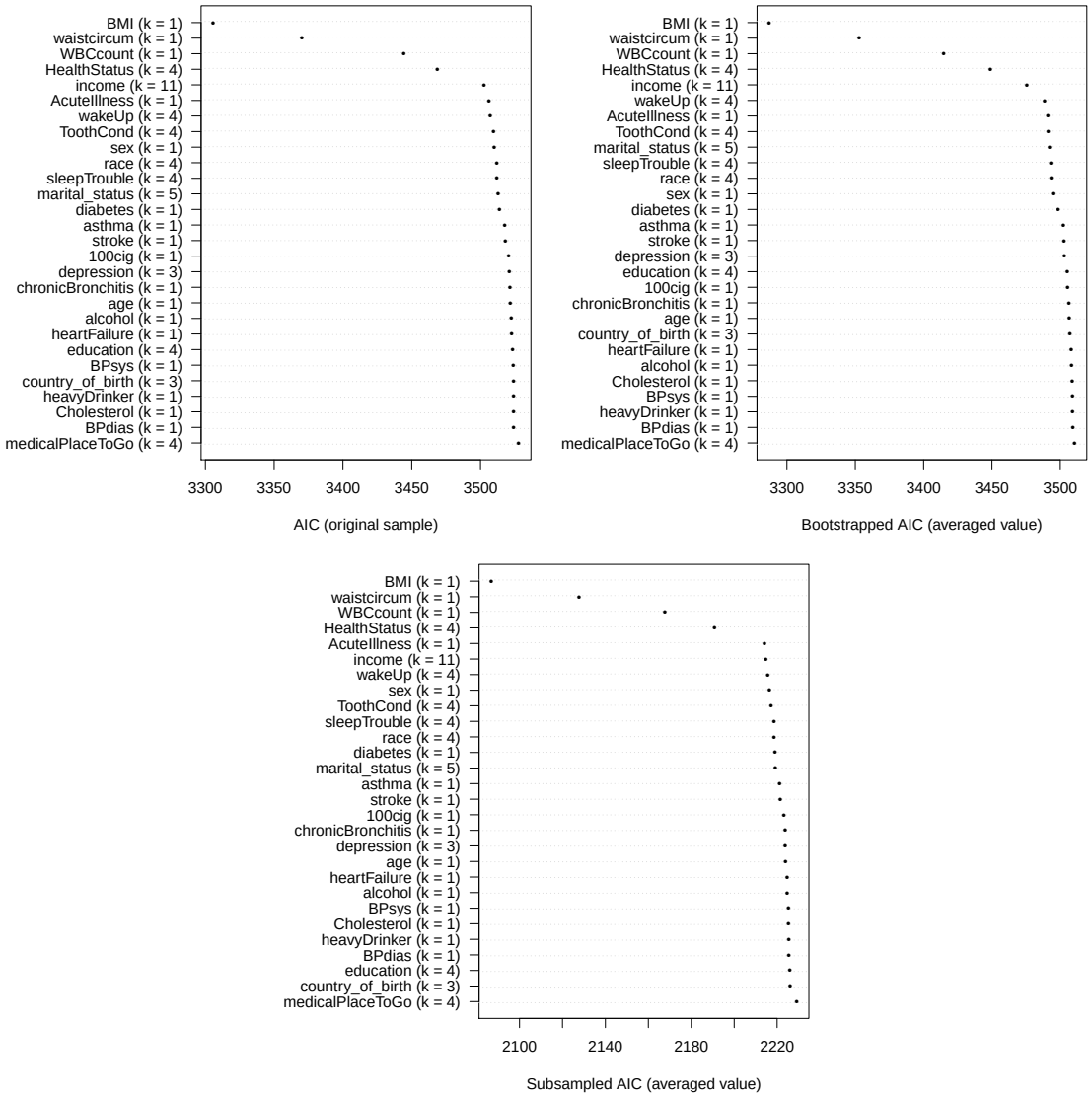

Figure 9: AIC values (in ascending order from top to bottom) obtained for the 28 models (each including exactly one covariate). The parameter k denotes the number of parameters included in the model for the respective variable. Upper left: AIC values derived on the original NHANES sample. Upper right: AIC values obtained from averaging over $B = 10000$ bootstrapped AIC values. Lower: AIC values obtained from averaging over $B = 10000$ AIC values computed based on subsamples.

Figure 10 shows the difference between the AIC values computed on the original NHANES sample and the average bootstapped AIC value. The difference seems to be bigger for models that include more parameters. Though all models have a systematically smaller bootstrapped AIC value, those models incorporating larger numbers of parameters have an exceedingly small AIC value, leading to the observed change in model ranking: more complex models have higher positions. There are three exceptions: the model featuring WBCcount, that for BMI and that for waistcircum. Note that these models are the models with the best model fit according to the AIC.

Results were also obtained when using subsamples instead of bootstrap samples. Since subsam-

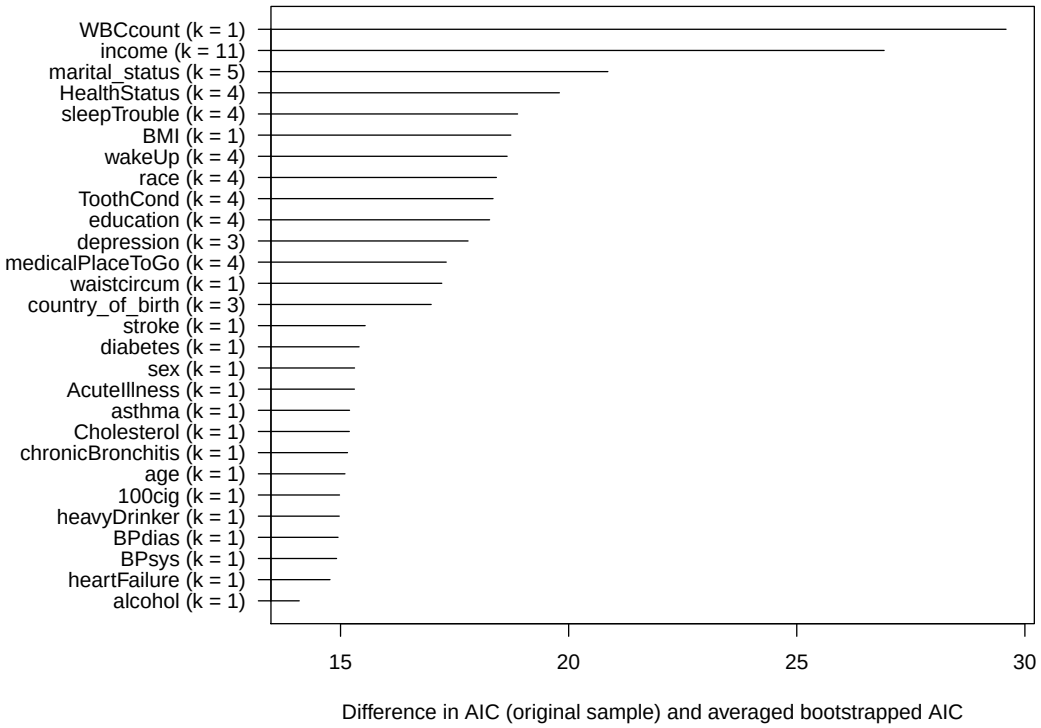

Figure 10: Difference between the AIC value computed on the original NHANES sample and the AIC value obtained from averaging over $B = 10000$ bootstrapped AIC values for 28 univariate models. Parameter k denotes the number of parameters estimated for the respective variable in the univariate linear model.

ples contain fewer observations, AIC values obtained for models on subsamples are not comparable to those obtained for the original sample. However, it is interesting to explore if the decision for or against a model is different when the AIC is computed on subsamples instead of the original sample. This can again be seen when sorting the models according to their AIC values (Figure 9, lower panel).

Indeed there are some characteristic changes in the ordering of the models according to the average AIC obtained from subsamples. But in contrast to the bootstrap, it seems as if more complex models (in terms of included parameters) are rather disfavored (see also Table A3 in the appendix). This can be explained as follows: From the definition of the AIC in Eq. (6) we can see that the AIC is dominated by the penalty term $2p$ (which penalizes the complexity of the model) if the first term $-2\log(L)$ is small, or equivalently, if the likelihood is large. Conversely, the AIC is dominated by the first term, $-2\log(L)$ (which is a measure of the model fit to the data), if the likelihood is small. The likelihood, as a product of n probabilities, becomes automatically smaller with increasing n. As a consequence the likelihood derived from a subsample is smaller than the likelihood of the original sample.

From these considerations it is clear that for subsamples the AIC is more driven by the penalty term than for the original sample, which leads to the observed phenomenon that more complex models are more disfavored in subsamples than in the original sample.

To conclude, AICs obtained from subsamples and original samples do not lead to the same conclusion regarding the choice of optimal models as well.

4.5 Application of Bootstrapped Information Criteria for Model Selection

In this section we investigate whether there is a preference for more complex models (in terms of included parameters) when constructing models based on bootstrap samples in the special context of gradient boosting (Friedman; 2001; Bühlmann et al.; 2007). Gradient boosting has become a popular method in biometrical applications to find sparse models by only making use of relevant predictor variables, which greatly facilitates model interpretation. Briefly, the idea of gradient boosting algorithms is to combine weak learners in an iterative fashion to obtain a strong learner with high prediction accuracy. The prediction accuracy depends highly on the number of iterations, also called the number of boosting steps. With too many boosting steps, many weak learners are constructed and the resulting strong learner might be overfit to the data and thus have poor prediction accuracy on new data. If the number of boosting steps is too small, the number of weak learners might be too small to appropriately model the relationship between the covariates and the response. Thus the number of boosting steps has to be carefully chosen, for example through application of information criteria or internal cross-validation. For more details on gradient boosting we refer the reader to the literature.

In the following analysis we apply the gradient boosting method firstly to the original NHANES sample and then to bootstrap samples. Again, the CRP level is the response variable. We use the AIC for choosing the number of boosting steps. Note that in contrast to the earlier analysis we now model the association between CRP level and the covariates in a multivariate fashion.

For the original NHANES sample, the number of boosting steps for the model with the smallest AIC was 309, the result being a model of 42 parameters (not including the intercept term). When performing tuning parameter selection on bootstrap samples we obtained systematically larger values for the number of boosting steps: in almost all (978 of $B = 1000$) bootstrap samples the chosen number of boosting steps was greater than 309 (see left boxplot in Figure 11). The mean number of boosting steps in bootstrap samples was 468. The resulting models included a larger number of parameters on average: the average number was 44.3, two parameters more than the model which was obtained for the original NHANES sample. The left panel in Figure 12 shows the relative frequency of models with a specific number of parameters. In 68.3% of the bootstrap samples the model included more than 42 parameters, in 24.7% the number of parameters was lower and in 7% the models included exactly 42 parameters.

We performed the same calculations using subsamples instead of bootstrap samples. As one would expect, sparser models were selected (on average 34.7 parameters) than for the original sample or bootstrap samples (right panel in Figure 12). Or equivalently, for subsamples a smaller number of boosting steps (254 on average) was chosen, seen in Figure 11 (right boxplot).

We also evaluated the models with respect to their predictive accuracy, using the observations that were not drawn into the bootstrap and subsample, respectively. Though models constructed on subsamples included fewer parameters, their predictive accuracy was comparable to the accuracy of models constructed on bootstrap samples: on average, even a marginally smaller mean squared error was obtained for models fit on subsamples (0.00075 compared to 0.00085 when using bootstrap samples), which suggests that the additional parameters in the models from bootstrap samples do not have any additional predictive value.

The overcomplexity induced by the bootstrap was more deeply investigated through simulation studies. Here the AIC was again used to determine the optimal number of boosting steps. We

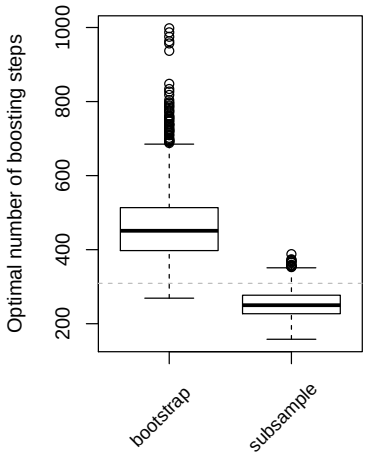

Figure 11: Optimal number of boosting steps selected via AIC in $B = 1000$ bootstrap samples and subsamples of the NHANES data. The dashed horizontal line indicates the chosen number of boosting steps in the original NHANES data.

now only show the results on the number of boosting steps since this number is directly related to the number of parameters included in the model and thus can be seen as a measure for the complexity. The data generating process is the same as that described by Binder and Schumacher (2008) for their simulation study on binary response gradient boosting. Data was simulated for the uncorrelated setting, where $p \in \{200, 1000, 5000\}$ predictors were independently drawn from a standard normal distribution for $n = 100$ observations. The covariate effects are defined as follows:

$$\beta_j = \begin{cases} c_e, & \text{if } j \cdot 200/p \in \{1, 3, 5, 7, 9\} \\ -c_e, & \text{if } j \cdot 200/p \in \{2, 4, 5, 6, 10\} \\ 0, & \text{otherwise} \end{cases}$$

where $c_e = 1$ (setting with weak effects) and $c_e = 2$ (setting with medium effects), as per Binder and Schumacher (2008). The binary response value for an observation i with covariates $\boldsymbol{x}_i$ was simulated from a Bernoulli distribution with success probability $\pi_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})/(1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}))$, with $\boldsymbol{\beta}^\top = (\beta_1, \ldots, \beta_p)$. We determined the optimal number of boosting steps on the original data, one bootstrap sample and one subsample. This was repeated 1000 times.

Figure 13 shows the optimal number of boosting steps for the setting with weak effects. The results for the setting with moderate effects are comparable and are thus not shown. The results of the simulation studies support our previous findings that a higher number of boosting steps, or equivalently, a higher complexity of gradient boosting models, is chosen when performing tuning parameter selection on bootstrap samples. However, the amount of overcomplexity induced by the bootstrap seems to be negligible in these studies, as seen by the only marginally higher number of boosting steps chosen. The discrepancy in selected boosting steps for original samples and for subsamples was much more pronounced than for original samples and bootstrap samples: a substantially smaller number of boosting steps was selected when using subsamples.

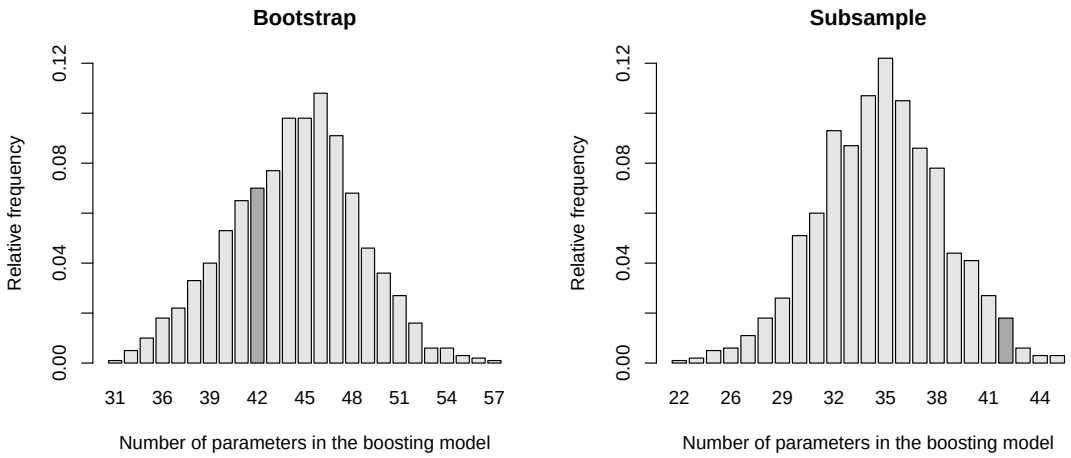

Figure 12: Relative frequency of boosting models (out of $B = 1000$) fitted on bootstrap samples (left) and on subsamples (right) with specified number of parameters (not including the intercept term). The dark gray bars indicate the number of parameters in the model that was fit on the original NHANES sample.

5 Discussion and Outlook

Bootstrap procedures are widely used in biometry to solve problems that are difficult to address using asymptotic theory. They can be applied for example to assess the variance of a statistic, a quantile of interest or for significance testing by resampling from the null hypothesis. Often hypothesis testing is performed by the use of the bootstrap. With bootstrap tests a p-value can be derived based on several bootstrap samples; these tests have been deeply investigated in the past. Bootstrapping p-values, however, is different from the famous bootstrap hypothesis tests and is the process in which each p-value is derived based on one single bootstrap sample. It has been shown that, when applying hypothesis testing on a bootstrap sample as if it were the original sample, the type I error is increased (Bollen and Stine; 1992; Strobl et al.; 2007). Accordingly, p-values obtained from bootstrap samples in this way should not be interpreted as standard p-values. Although bootstrapping p-values as considered in our paper is less popular than the classical bootstrap hypothesis tests, important approaches which make use of bootstrapped p-values have been suggested in the literature and we believe that similar such procedures will continue to be proposed in the future. There is a need for studies like ours investigating their potential pitfalls.

Similar problems apply to approaches which make use of bootstrapped information criteria like the AIC or BIC. Such approaches have been proposed and are in use for model building in biometrical applications. Evidence that information criteria computed from bootstrap samples depart from that of original samples has been given, for example, by Steck and Jaakkola (2003) and Wagenmakers et al. (2004). However, there is a lack of studies that explore whether the systematic deviation in bootstrapped information criteria affects the reliability of the results obtained by a researcher.

In this article we applied selected bootstrap-based approaches (making either use of bootstrapped p-values or bootstrapped information criteria) on a large real data set from a population-based study to investigate whether results are affected by the systematic deviation in bootstrapped

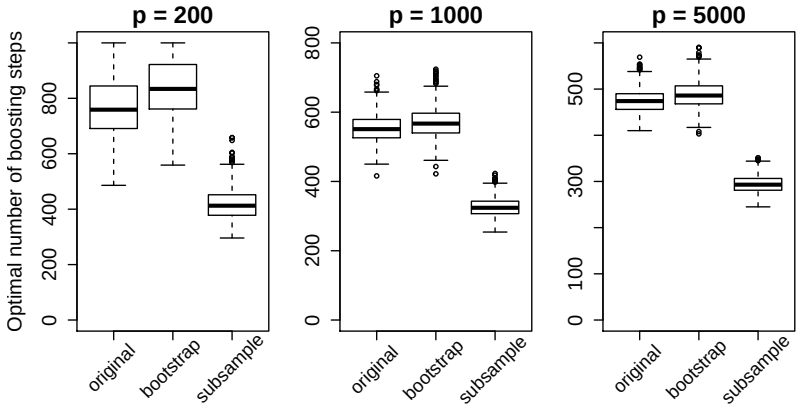

Figure 13: Optimal number of boosting steps selected via AIC for binary response gradient boosting in 1000 original samples, bootstrap samples and subsamples for the setting with weak effects ($c_e = 1$).

p-values and information criteria. When univariately testing the association between the level of high-sensitive C-reactive protein and various factors, we observed that bootstrapped p-values are often considerably smaller than p-values that are obtained for the original data. Also seen in our studies was that making decisions based on bootstrapped p-values results in increased number of false positive results. Further, the variability of bootstrapped p-values was shown not to reflect the variability of p-values when repeating the same experiment several times, thus making the reliability of the approach suggested by Boos and Stefanski (2011) questionable.

We also observed a bias in bootstrapped information criteria when these were compared to information criteria that were derived from the original sample. In our studies this led to a preferential selection of models which included more parameters, since these models systematically had smaller bootstrapped AIC values. Further, bootstrapped AIC values are sometimes used in the context of gradient boosting models. Here the tuning parameter selection (via AIC) and model fitting is performed based on a bootstrap sample while the remaining observations that were not drawn into the bootstrap sample are used for evaluating the model. In our application on real data we observed higher values for the tuning parameter for bootstrap samples. This led to more complex boosting models (i.e., more parameters) than the model fit on the original sample. These results are in line with those reported by Steck and Jaakkola (2003) in the context of graphical models who show that more complex models (in terms of included parameters), have actually too high a likelihood, or equivalently, too small an AIC value, when fit on bootstrap samples. Thus when using the AIC to select a model that was built from a bootstrap sample, one gives preference to more complex models which would possibly not be selected had the original sample been used.

We also investigated the use of subsampling as a promising alternative strategy to circumvent biases induced by the bootstrap. The properties of subsampling have been theoretically investigated in the literature; it has been shown that subsampling has desirable properties even in situations where the bootstrap fails. A recent approach to stability selection based on subsampling was introduced by Meinshausen and Bühlmann (2010). Their studies impressively show that subsampling is a powerful tool in investigating the stability of models, such as penalized likelihood models and graphical models. Further Strobl et al. (2007) proposed the use of subsampling instead of bootstrapping in the context of random forests to circumvent the problem

of preferential selection of certain types of predictors for a split. However, our results show that subsampling should not be regarded as an universally applicable alternative to the bootstrap. For selecting the optimal number of boosting steps via information criteria, for example, with the subsampling procedure a smaller number of boosting steps is selected, which might lead to too sparse models. If the aim is to investigate the distribution of model complexity parameters, the subsampling procedure is thus not recommended; prediction performance might similarly be affected. For investigating the variability of p-values, subsampling is again not appropriate, if more than type I error control is wanted. Our investigations make it clear that subsampling is not a reliable alternative to the bootstrap for all types of applications, even if it has shown important advantages in some situations (Strobl et al.; 2007; De Bin et al.; 2014).

Applied researchers should be careful when using approaches to problems in which hypothesis tests or information criteria are computed based on a bootstrap sample. If no investigations exist that indicate the reliability of a bootstrap approach, simulation studies are a helpful investigative tool. It is important to keep in mind that one cannot directly apply any procedure to bootstrap samples as if they were the original sample. It is advisable for methodologists to check the validity of their proposed bootstrap approaches by using simulation studies, and then comparing these results to those obtained when using original samples from the true underlying distribution. In this way unexpected results can be easily discovered and adjustments may be made.

Acknowledgements

SJ was financed by grant BO3139/2-2 to ALB (DFG Einzelförderung). The authors thank Rory Wilson for helpful comments.

Supplementary Material

R code is available at `http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/070_drittmittel/janitza/index.html` for full reproducibility of our studies.

Appendix

A.1 Data Description

Abbreviation	Interview question/description	Categories/units
race	Recode of reported race and ethnicity information	Mexican American Other Hispanic Non-Hispanic White Non-Hispanic Black Other Race - Including Multi-Racial
country_of_birth	In what country (were you/was SP) born?	50 US States or Washington, DC Mexico Other Spanish Speaking Country Other Non-Spanish Speaking Country
education	What is the highest grade or level of school (you have/SP has) completed or the highest degree (you have/she/he has) received?	less than 9th up to 11th high school some college graduate
marital_status	Marital staus	married widowed divorced separated never married living with partner
HealthStatus	Would you say (your/SP's) health in general is . . .	excellent very good good fair poor
depression	Over the last 2 weeks, how often have you been bothered by the following problems: little interest or pleasure in doing things? Would you say . . .	not at all several days over half the days nearly every day
ToothCond	Now I have some questions about the condition of your teeth and gums. How would you describe the condition of (your/SP?s) teeth? Would you say . . .	excellent very good good fair poor
sleepTrouble	In the past month, how often did (you/SP) have trouble falling asleep?	never rarely sometimes often almost always
wakeUp	In the past month, how often did (you/SP) wake up during the night and had trouble getting back to sleep?	never rarely sometimes often almost always
medicalPlaceToGo	What kind of place (do you/does SP) go to most often: is it a clinic, doctor's office, emergency room, or some other place?	clinic

Continued on next page

Continued from previous page

		doctor's office hospital emergency hospital outpatient other
income	Total household income (reported as a range value in dollars)	under $5k $5k - under $10k $10k - under $15k $15k - under $20k $20k - under $25k $25k - under $35k $35k - under $45k $45k - under $55k $55k - under $65k $65k - under $75k $75k - under $100k over $100k
AcuteIllness	Did (you/SP) have a head cold or chest cold that started during the last 30 days? *or* Did (you/SP) have flu, pneumonia, or ear infections that started during those 30 days? *or* Did (you/SP) have a stomach or intestinal illness with vomiting or diarrhea that started during those 30 days?	no yes
100cig	(Have you/Has SP) smoked at least 100 cigarettes in (your/his/her) entire life?	yes no
diabetes	(Have you/Has SP) ever been told by a doctor or health professional that (you have/(he/she/SP) has) diabetes or sugar diabetes?	yes no
asthma	Has a doctor or other health professional ever told (you/SP) that (you/she/he) have/has asthma?	yes no
heartFailure	Has a doctor or other health professional ever told (you/SP) that (you/she/he) had congestive heart failure?	yes no
stroke	Has a doctor or other health professional ever told (you/SP) that (you/she/he) had a stroke?	yes no
chronicBronchitis	Has a doctor or other health professional ever told (you/SP) that (you/she/he) had chronic bronchitis?	yes no
heavyDrinker	Was there ever a time or times in (your/SP's) life when (you/he/she) drank 5 or more drinks of any kind of alcoholic beverage almost every day?	yes no
waistcircum	circumvence of waist	cm
Cholesterol	cholesterol level	md/dl
WBCcount	white blood cell count	1k cells/μl
BPsys	systolic blood pressure	mmHg
BPdias	diastolic blood pressure	mmHg
age	age	years
BMI	body mass index	kg/m^2
alcohol	alcohol consume	units

Table A1: Variables and corresponding interview question or description for the considered NHANES data.

A.2 Empirical Studies on the Distribution of a Bootstrapped Z-test Statistic

In this section we present empirical results that complement the results presented in Section 3.1.1. The notation is the same as introduced in 3.1.1. For computing Z and Z^* we draw $n = 1000$ independent observations from the standard normal distribution. We then draw a bootstrap sample out of this original sample and compute the test statistic for a Z-test with null hypothesis $H_0 : \mu = 0$ from both original and bootstrap samples. This procedure is repeated 500000 times, yielding 500000 values of both Z and Z^*. Figure A1 shows the resulting empirical density functions of Z and Z^*. As expected from theory the distribution of the test statistic Z coincides with the standard normal distribution: the two lines in Figure A1 completely overlap. The distribution of the test statistic Z^* in contrast systematically deviates from the standard normal distribution. There is a remarkable difference in variances of the test statistics Z and Z^* while the expected values seem to be equal. The empirical expectation of Z and Z^* are both close to zero with values -0.0010 and 0.0018, respectively. In contrast, the empirical variance of Z^* is, at 2.0011, larger by a factor of 2 than the variance of Z, which is, at 1.0009, very close to the variance of the standard normal distribution.

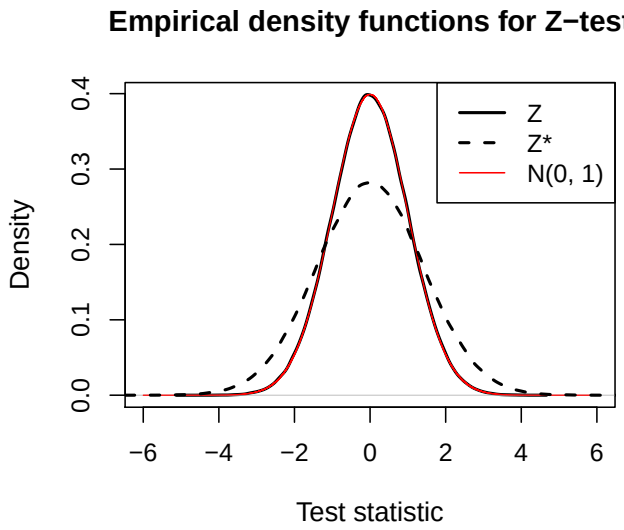

Figure A1: Empirical density functions for test statistics Z (solid black line) and Z^* (dashed black line) of the Z-test. The density of the standard normal distribution is indicated by the red line.

A.3 Additional Results of the Real Data Application

Scale	Variable	Original rank	Bootstrap rank	(diff.)	Subsample rank	(diff.)
metric or $m = 2$	diabetes	3	5	(−2)	2	(+1)
	asthma	4	7	(−3)	4	0
	heartFailure	6	8	(−2)	5	(+1)
	AcuteIllness	8	11	(−3)	8	0
	BPsys	10	12	(−2)	10	0
	age	11	13	(−2)	11	0
	alcohol	12	19	(−7)	12	0
	BPdias	13	18	(−5)	13	0
	stroke	14	21	(−7)	14	0
	heavyDrinker	15	22	(−7)	16	(−1)
	sex	16	20	(−4)	15	(+1)
	chronicBronchitis	18	17	(+1)	18	0
	WBCcount	19	28	(−9)	24	(−5)
	waistcircum	22	23	(−1)	19	(+3)
	BMI	23	26	(−3)	23	0
	Cholesterol	25	24	(+1)	20	(+5)
	100cig	27	25	(+2)	22	5
$m = 4$	depression	20	16	(+4)	25	(−5)
	country_of_birth	28	27	(+1)	28	0
$m = 5$	sleepTrouble	1	2	(−1)	1	0
	medicalPlaceToGo	5	4	(+1)	6	(−1)
	wakeUp	9	6	(+3)	9	0
	race	17	9	(+8)	17	0
	education	21	10	(+11)	21	0
	HealthStatus	24	14	(+10)	26	(−2)
	ToothCond	26	15	(+11)	27	(−1)
$m = 6$	marital_status	7	3	(+4)	7	0
$m = 12$	income	2	1	(+1)	3	(−1)

Table A2: Variable ranking for the first of the 1000 modified NHANES data sets (modification consisted of permuting the response variable). Variable rankings are determined by p-values obtained for the original sample ("Original rank"), by the median bootstrapped p-value ("Bootstrap rank"), and by the median p-value from subsamples ("Subsample rank"). The difference to the "Original rank" is given in brackets for each variable. The parameter m denotes the number of levels for the categorical predictor variables.

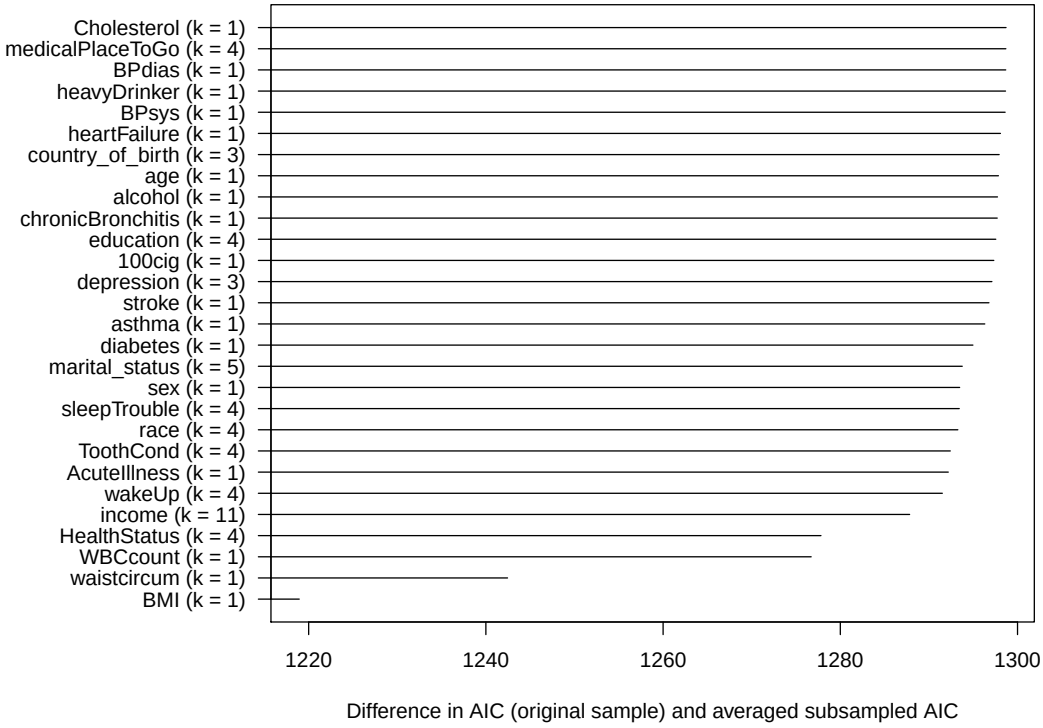

Figure A2: Difference between the AIC value computed on the original NHANES sample and the AIC value obtained by averaging $B = 10000$ AIC values computed based on subsamples, for 28 univariate models. Parameter k denotes the number of parameters to be estimated for the respective variable in the univariate linear model.

Model complexity	Included variable	Position (original sample)	Bootstrap position	(diff.)	Subsample position	(diff.)
$k = 1$	BMI	1	1	(0)	1	(0)
	waistcircum	2	2	(0)	2	(0)
	WBCcount	3	3	(0)	3	(0)
	AcuteIllness	6	7	(−1)	5	(+1)
	sex	9	12	(−3)	8	(+1)
	diabetes	13	13	(0)	12	(+1)
	asthma	14	14	(0)	14	(0)
	stroke	15	15	(0)	15	(0)
	100cig	16	18	(−2)	16	(0)
	chronicBronchitis	18	19	(−1)	17	(+1)
	age	19	20	(−1)	19	(0)
	alcohol	20	23	(−3)	21	(−1)
	heartFailure	21	22	(−1)	20	(+1)
	BPsys	23	25	(−2)	22	(+1)
	heavyDrinker	25	26	(−1)	24	(+1)
	Cholesterol	26	24	(+2)	23	(+3)
	BPdias	27	27	(0)	25	(+2)
$k = 3$	depression	17	16	(+1)	18	(−1)
	country_of_birth	24	21	(+3)	27	(−3)
$k = 4$	HealthStatus	4	4	(0)	4	(0)
	wakeUp	7	6	(+1)	7	(0)
	ToothCond	8	8	(0)	9	(−1)
	race	10	11	(−1)	11	(−1)
	sleepTrouble	11	10	(+1)	10	(+1)
	education	22	17	(+5)	26	(−4)
$m = 5$	medicalPlaceToGo	28	28	(0)	28	(0)
	marital_status	12	9	(+3)	13	(−1)
$m = 11$	income	5	5	(0)	6	(−1)

Table A3: Model rank by AIC computed for the original sample ("Position (original sample)"), by the average bootstrapped AIC ("Bootstrap position"), and by the average subsampled AIC ("Subsample position"). The difference to "Position (original sample)" is given in brackets for each model. The parameter k denotes the number of included parameters in a model and thus is a measure of model complexity.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *in* B. Petrov and F. Caski (eds), *Second International Symposium on Information Theory*.

Altman, D. G. and Andersen, P. K. (1989). Bootstrap investigation of the stability of a Cox regression model, *Statistics in Medicine* **8**: 771–783.

Bickel, P. J. and Sakov, A. (2005). On the choice of m in the m out of n bootstrap and its application to confidence bounds for extreme percentiles, *Unpublished manuscript* .

Binder, H. and Schumacher, M. (2008). Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples, *Statistical Applications in Genetics and Molecular Biology* **7**: 1.

Black, S., Kushner, I. and Samols, D. (2004). C-reactive protein, *Journal of Biological Chemistry* **279**: 48487–48490.

Bollen, K. A. and Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models, *Sociological Methods & Research* **21**: 205–229.

Boos, D. D. and Stefanski, L. A. (2011). P-value precision and reproducibility, *The American Statistician* **65**: 213–221.

Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997). Model selection: an integral part of inference, *Biometrics* **53**: 603–618.

Bühlmann, P., Hothorn, T. et al. (2007). Boosting algorithms: Regularization, prediction and model fitting, *Statistical Science* **22**: 477–505.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multi-model inference: a practical information-theoretic approach*, Springer, New York.

Chen, C.-H. and George, S. L. (1985). The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model, *Statistics in Medicine* **4**: 39–46.

Chernick, M. R. (2011). *Bootstrap methods: A guide for practitioners and researchers*, 2 edn, New York, John Wiley & Sons.

Davison, A. C. (1997). *Bootstrap methods and their application*, Vol. 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press.

Davison, A. C., Hinkley, D. V. and Young, G. A. (2003). Recent developments in bootstrap methodology, *Statistical Science* **18**: 141–157.

De Bin, R., Janitza, S., Sauerbrei, W. and Boulesteix, A.-L. (2014). Subsampling versus bootstrap in resampling-based model selection for multivariable regression, *Technical Report 171*, Department of Statistics, University of Munich.

Efron, B. (1979). Bootstrap methods: another look at the jackknife, *The Annals of Statistics* **7**: 1–26.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*, Vol. 57, CRC Press.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine, *Annals of Statistics* **29**: 1189–1232.

Good, P. I. (2005). *Permutation, parametric and bootstrap tests of hypotheses*, 3 edn, New York, Springer.

Hartigan, J. A. (1969). Using subsample values as typical values, *Journal of the American Statistical Association* **64**: 1303–1317.

Manly, B. F. (2006). *Randomization, bootstrap and Monte Carlo methods in biology*, 3 edn, Florida, CRC Press.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**: 417–473.

Mukherjee, S., Roberts, S. J., Sykacek, P. and Gurr, S. J. (2003). Gene ranking using bootstrapped p-values, *ACM SIGKDD Explorations Newsletter* **5**: 16–22.

National Center for Health Statistics (2012). NHANES 2007 to 2008 public data general release file documentation, http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/generaldoc_e.htm.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions, *The Annals of Statistics* **22**: 2031–2050.

Politis, D., Romano, J. and Wolf, M. (1999). *Subsampling*, Springer, New York.

Rospleszcz, S., Janitza, S. and Boulesteix, A.-L. (2014). Categorical variables with many categories are preferentially selected in model selection procedures for multivariable regression models on bootstrap samples, *Technical Report 164*, Department of Statistics, University of Munich.

Sauerbrei, W., Boulesteix, A.-L. and Binder, H. (2011). Stability investigations of multivariable regression models derived from low-and high-dimensional data, *Journal of Biopharmaceutical Statistics* **21**: 1206–1231.

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building: application to the Cox regression model, *Statistics in Medicine* **11**: 2093–2109.

Shao, J. and Wu, C. J. (1989). A general theory for jackknife variance estimation, *The Annals of Statistics* **17**: 1176–1197.

Steck, H. and Jaakkola, T. S. (2003). Bias-corrected bootstrap and model uncertainty, *Advances in Neural Information Processing Systems*, number 16.

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution, *BMC Bioinformatics* **8**: 25.

Wagenmakers, E.-J., Farrell, S. and Ratcliff, R. (2004). Naïve nonparametric bootstrap model weights are biased, *Biometrics* **60**: 281–283.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis, *The Annals of Statistics* **14**: 1261–1295.